International Face Performance Conference (IFPC) 2018, NIST, Gaithersburg MD, November 2018

DCNNs for Unconstrained Face Recognition

Carlos D. Castillo
carlos@umiacs.umd.edu
University of Maryland

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

What is this talk about?

- How to use Convolutional Neural Networks (CNNs) to recognize faces and …
- How a since September 2014 a University-only team built a deep learning based face recognition system that:
  - Can compute face representations that are comparable, searchable, indexable and clusterable.
  - Compares well to the performance of forensic face examiners in hard face matching tasks.
- Evaluating our system and understanding where it can improve.
- Summarize what we learned, and things we still don’t know.

CMU PIE Illumination Variation (2003)

CMU PIE (2003)

- Images from CMU PIE


CMU PIE (2003)

- Images from CMU PIE

Face Recognition 10 years ago

- Methods used:
  - Most if not all methods followed the same pattern
    - Compute gradients, optionally smooth the image before doing so
    - Pool the gradients
    - Train a classifier
  - Two common examples:
    - Histogram of Oriented Gradients/Dense Histogram Oriented Gradients
      - Compute gradients
      - Build histogram
    - LBP
      - Stand at each pixel x build a 1-0 descriptor based on each neighbor being > or <= than x
      - Build a histogram over each cell for each 1-0 pattern

Labeled Faces in the Wild (2008)

- Images from news stories from 2002-2004
- Images mostly of politicians (and other people in the news)
- Professional photographer
  - Probably selected as the "best image" from a set of 10 or 20
- Mostly frontal
- Still significantly more difficult than what was being done in face recognition back then

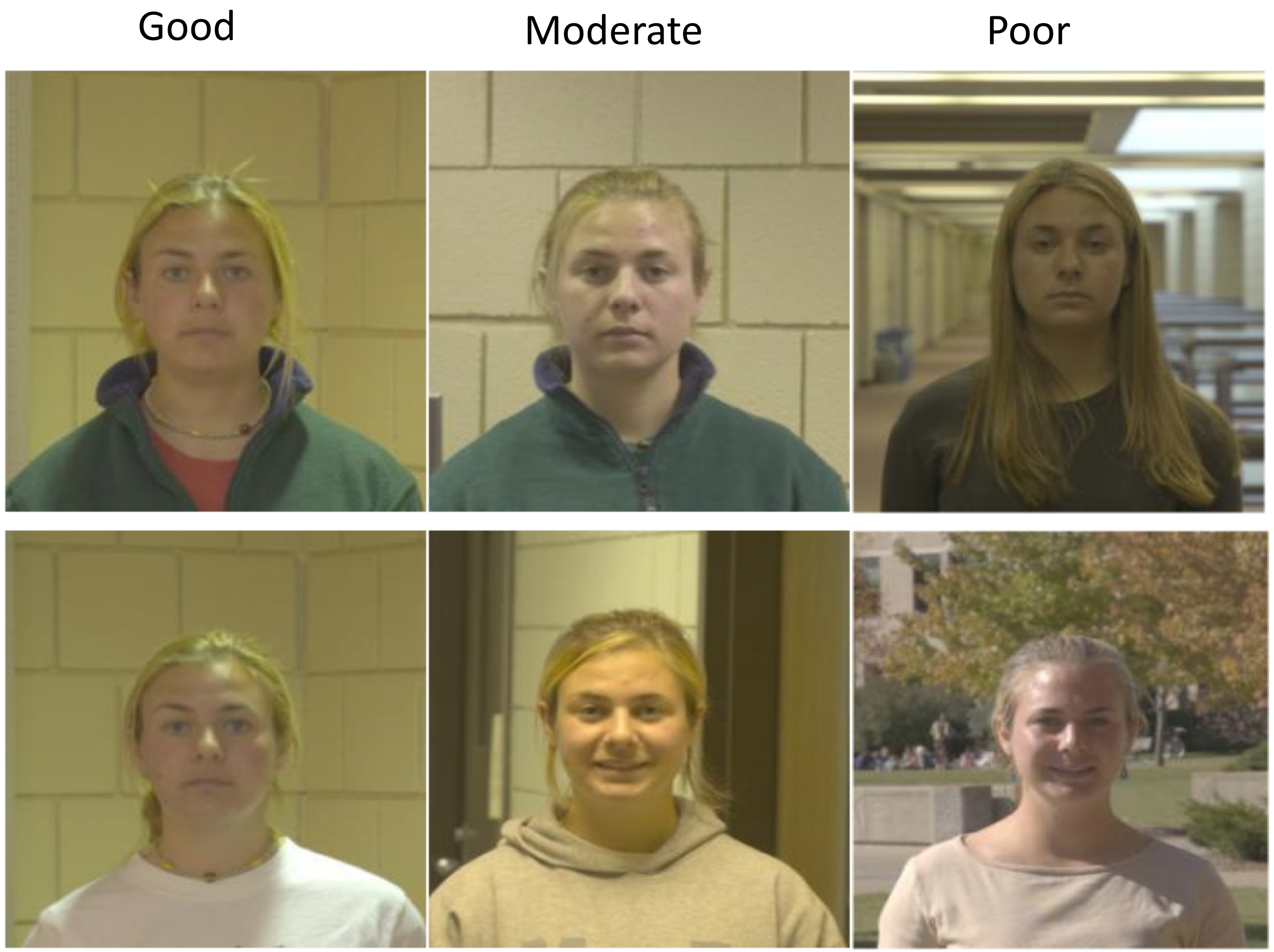
Labeled Faces in the Wild (2008)

- There were still some difficult images by today's standards:
  - Actors and actresses
  - Not frontal
  - Event photography




The Good, the Bad and the Ugly (2007-2012)

- Frontal images
- Acquired throughout an academic year
- Indoor and outdoor
- Carefully analyzed
  - We know which pairs are easy and which are hard

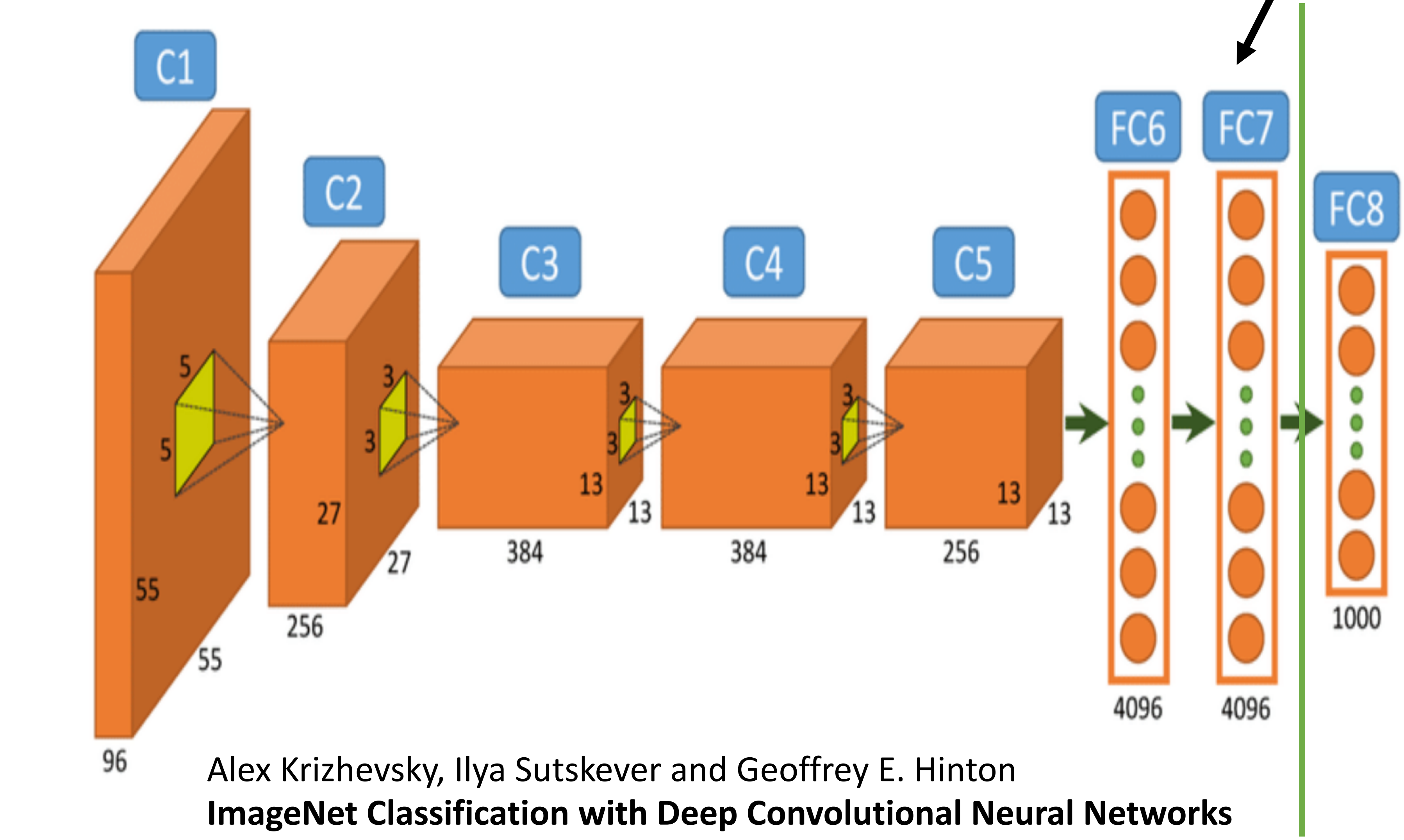


The Breakthrough

- Hindsight:
  - Gradients are just a simple type of convolution
  - Building a histogram is just a pooling operation
  - We can learn convolutions to describe faces from data
    - Need lots of data

AlexNet

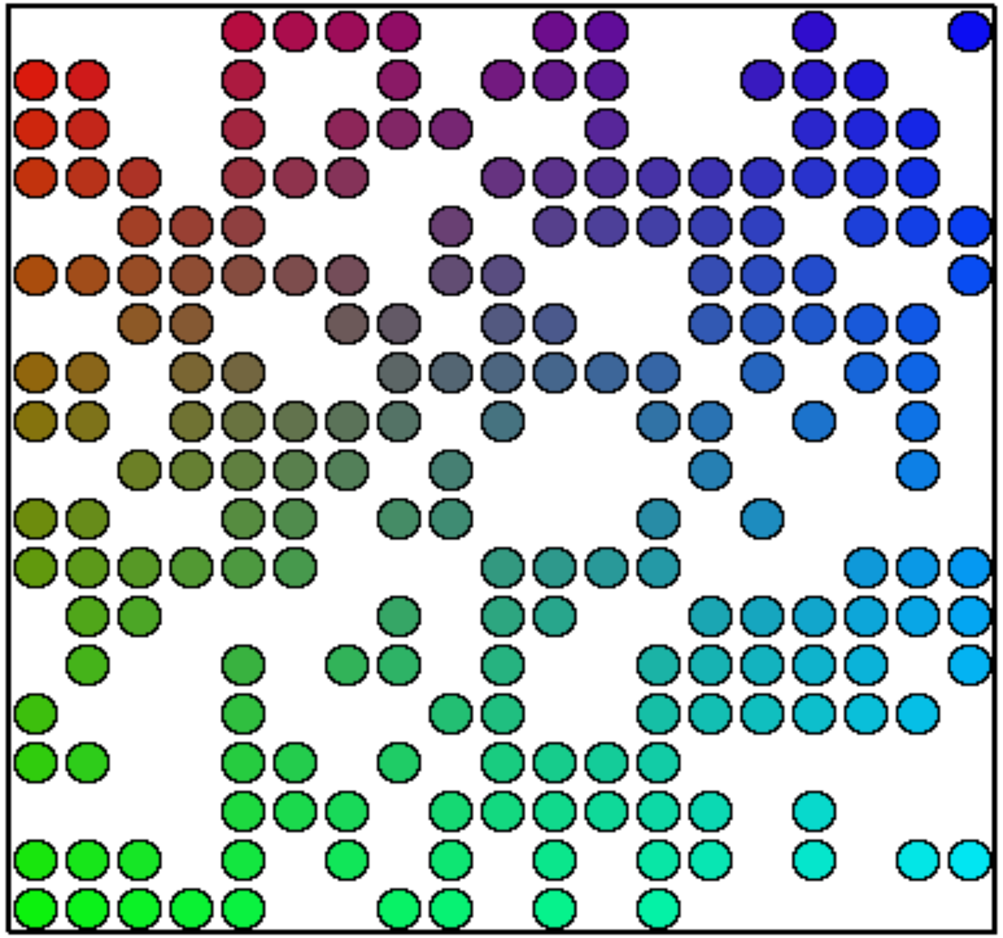
This is the face descriptor



Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton
ImageNet Classification with Deep Convolutional Neural Networks
NIPS 2012

Example: Marbles on a Table

- If the table is too small there won't be enough space
- If the table is too big you will be wasteful organizing the marbles
- But if the relationship of space to colors is just right, you will be able to organize the marbles in such a way that you will be able find a location even for colors you've never before



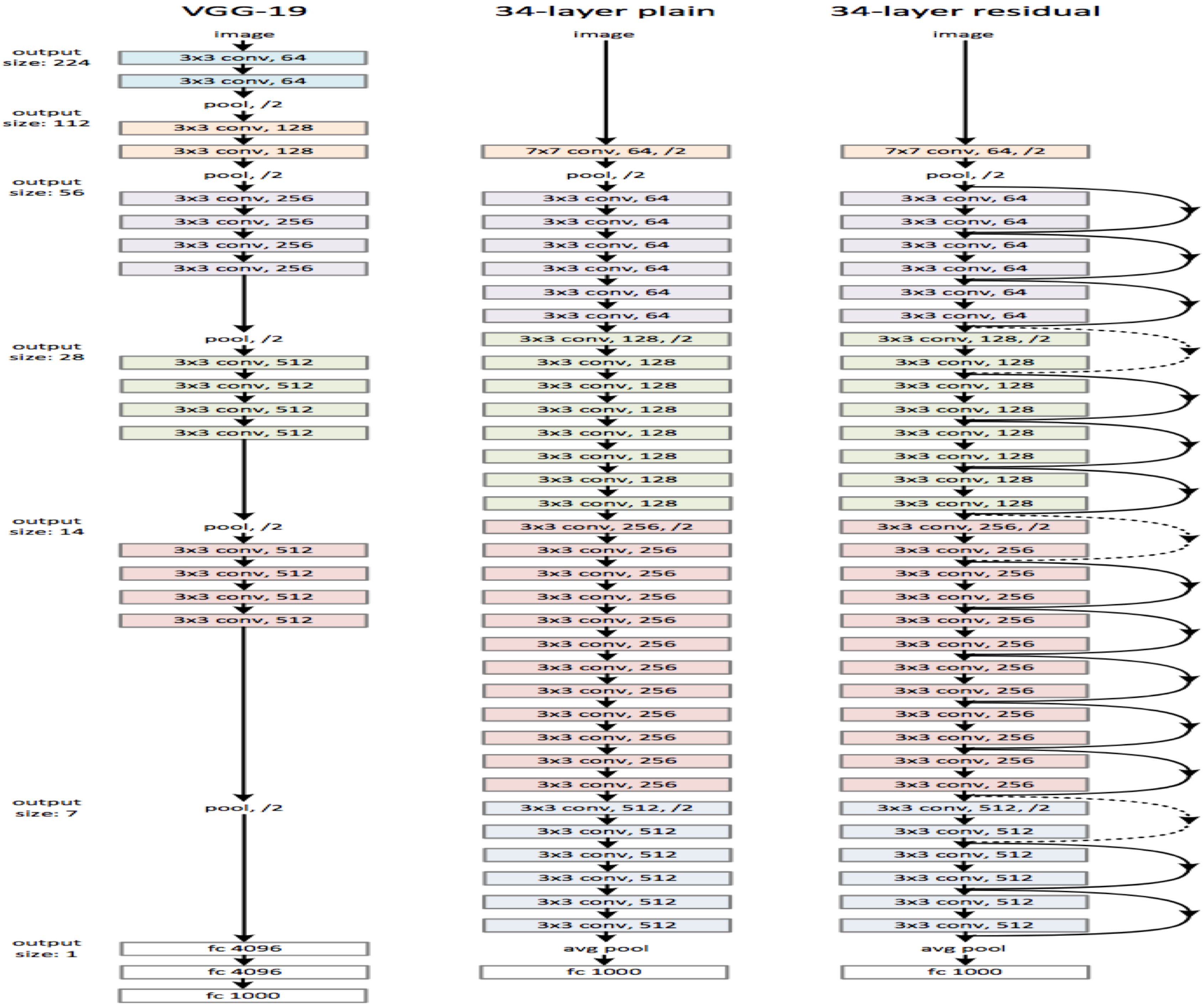
Changes in the last 10-15 years

- Back then:
  - Features were fixed – defined by the researcher and not obtained from the data or task.
  - We learned classifiers on those fixed features
- Now:
  - Features are learned from the data – learned by applying backpropagation on the data to classify it effectively.
  - Classifiers are trained on those learned features

Why didn’t this work before?

- New insights (2012-):
  - Disconnect the head (or decapitate), use the previous layer as features for recognition
    - High level visual features are pretty generic
  - ReLU:
    - Activations between layer add non-linearities
    - Sigmoid and tanh were used originally
  - GPU, faster CPUs, more storage
    - Everything is bigger and faster than 20 years ago
  - Data augmentation
    - Flip faces
    - Center cropping

Alice J. O’Toole, Carlos D. Castillo, Connor J. Parde, Matthew Q. Hill, Rama Chellappa
Face Space Representations in Deep Convolutional Neural Networks
Trends in Cognitive Sciences, 2018


VGG-19
34-layer plain
34-layer residual
output size: 224
output size: 112
output size: 56
output size: 28
output size: 14
output size: 7
output size: 1
image
3x3 conv, 64
3x3 conv, 64
pool, /2
3x3 conv, 128
3x3 conv, 128
pool, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
fc 4096
fc 4096
fc 1000
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000

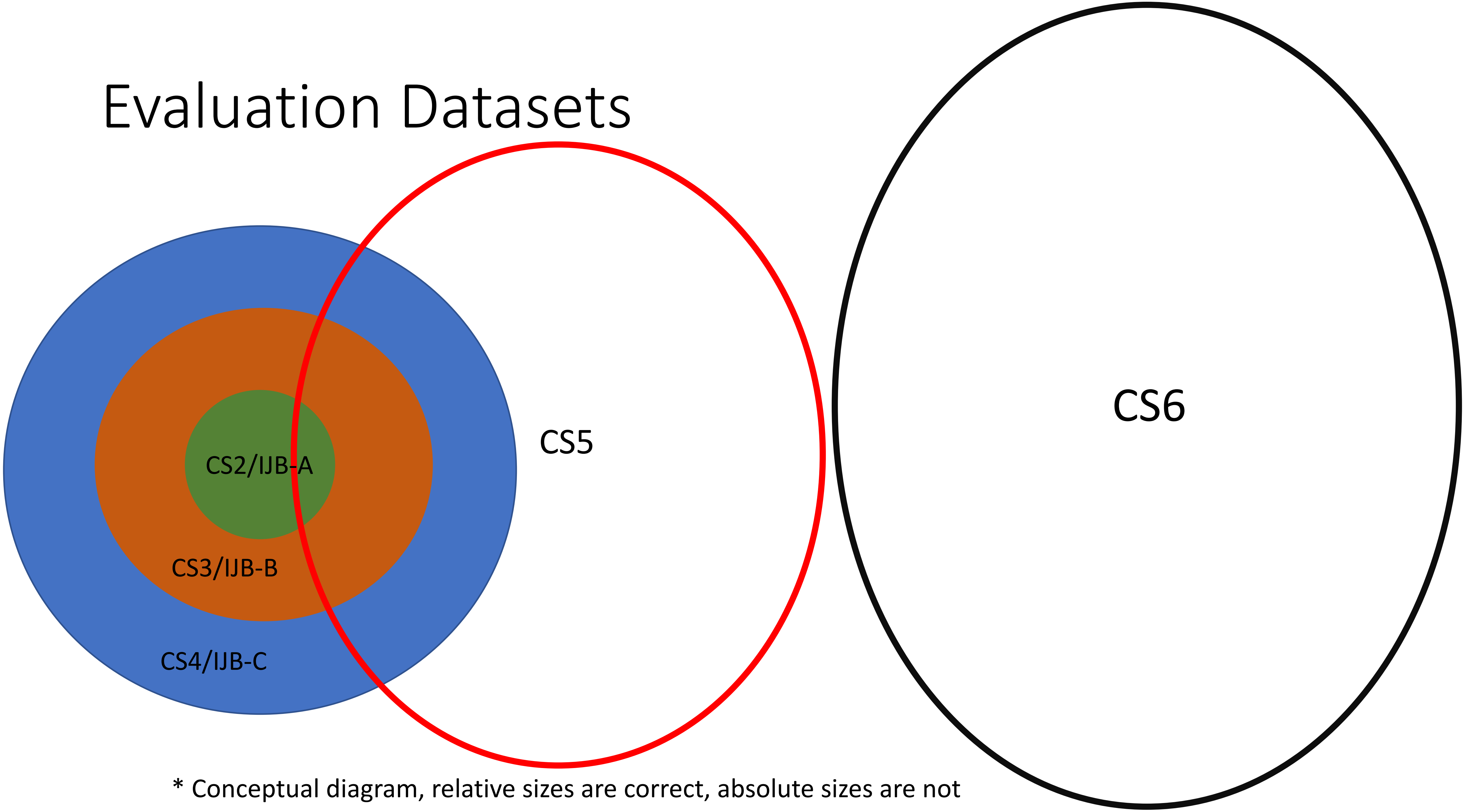
Training Datasets Used these Days

- Publicly Available Datasets:
  - MS1M: 100,000 individuals, 10 million images
  - CASIA: 10,575 individuals, 494K images
  - VGG: 2.6M images of 2622 subjects
  - VGG2: 3.31 million images of 9131 subjects
  - CelebA: 10,177 individuals, 202K images
  - UMDFaces: 8,277 individuals, 367K images
    - UMDFaces (videos): 3000 individuals, 3.7 million frames
- Companies:
  - Facebook: has a paper in which they train with a 10 million identity dataset with more than 80 million images
  - Google: more than 200 million faces

Evaluation Datasets

- IJB-A, IJB-B and IJB-C and CS2/3/4, CS 5 (huge gallery) and CS 6 (surveillance)
    - Really unconstrained
        - In different ways
    - Some images, some frames from videos
        - Cure for the photographic bias in LFW
    - Template based: a template is a set of images of the same person
    - Many tasks: verification, search, covariates, clustering, uncurated search, video probes.

Evaluation Datasets



* Conceptual diagram, relative sizes are correct, absolute sizes are not

Evaluation Tasks

- This a selection of the tasks on which we are evaluated:
  - Template to template verification experiment with 10^{12} evaluation pairs
  - 1:N template search with a gallery with 1.1 million templates
  - 1:N search of uncurated probes against a large gallery
  - Clustering of templates into identities
  - Video uncurated search
- Athletes and Events – Algorithms and Tasks

Our Approach: The Decathlete

Events:

- 100 m
- Long jump
- Shot put
- High jump
- 400 m
- 110 m
- Discus throw
- Pole vault
- Javelin throw
- 1500 m

UltraFace

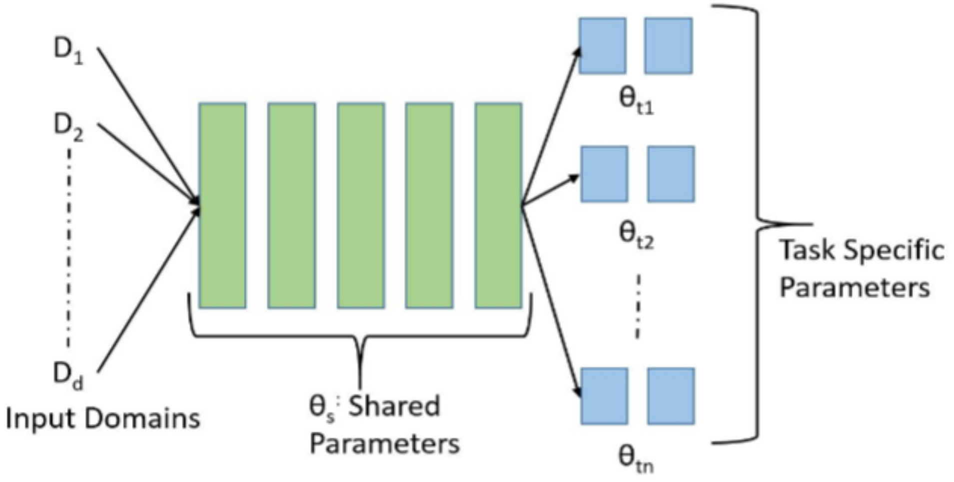
Rajeev Ranjan, Carlos D. Castillo, Rama Chellappa
An all-in-one convolutional neural network for face analysis
FG 2017

- First order of business:
  - Reliably be able to detect and obtain key points and estimate attributes on images like these:

Idea: explore performing all of these tasks in an all in one approach.

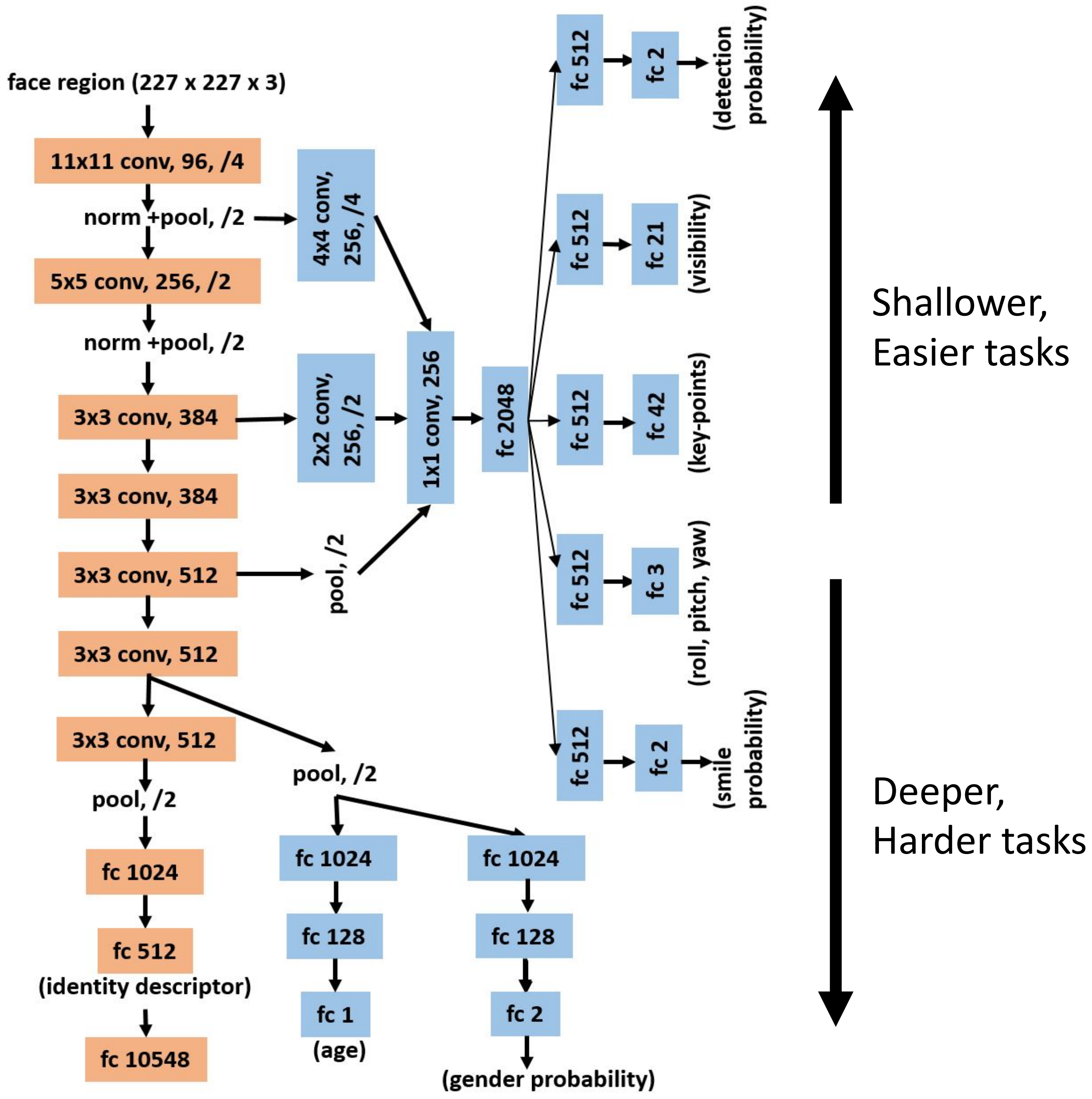
UltraFace

- Task-based
  - shrinks the solution space of θ_s such that the learned parameter vector is in consensus with all the tasks
- Domain-based
  - θ_s adapts to the complete set of domains instead of fitting to a task-specific domain



UltraFace Architecture

- Parameters initialized from face identification network
- Subject-independent tasks (face-detection, fiducials, pose, smile) share the lower layers of the network
- Subject-specific features are pooled from deeper layers of the network



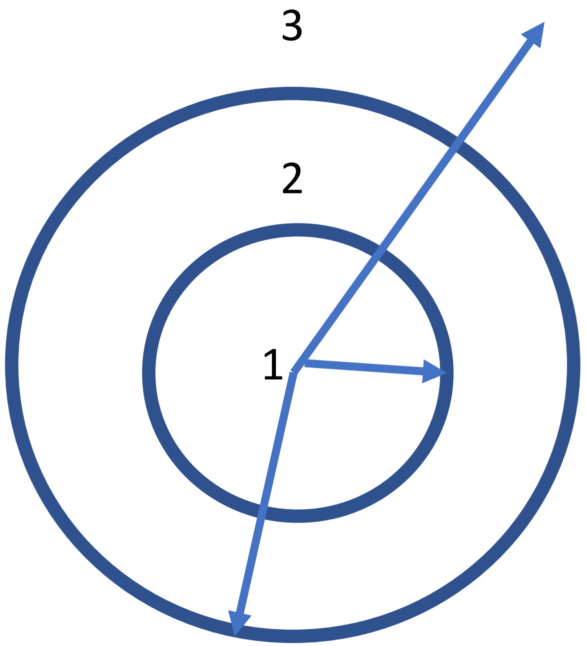
Examples of Data IJB-A (2015) /IJB-B (2017)

Descriptors and their L2 Norms

(1) Low norm descriptor:

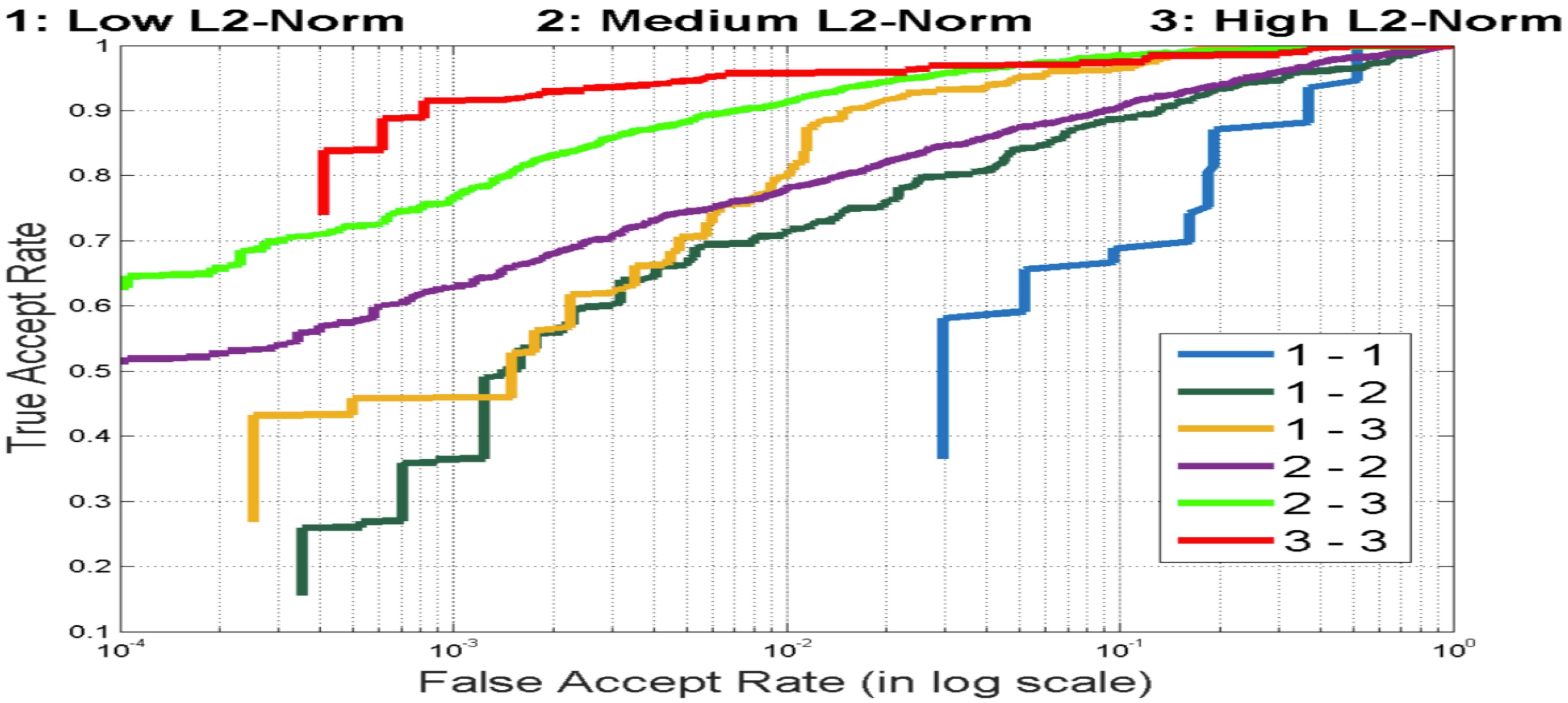
(2) Medium norm descriptor:

(3) High norm descriptor:



Connor J. Parde, Carlos D. Castillo, Matthew Q. Hill, Y Ivette Colon, Swami Sankaranarayanan, JC Chen, Alice J O'Toole
Face and Image Representation in Deep CNN Features
FG 2017

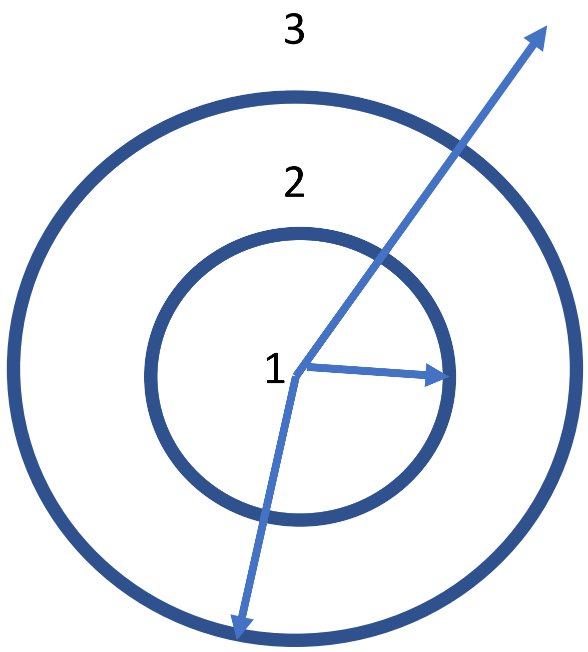
Performance Grouped by Norm of the Descriptors



recall

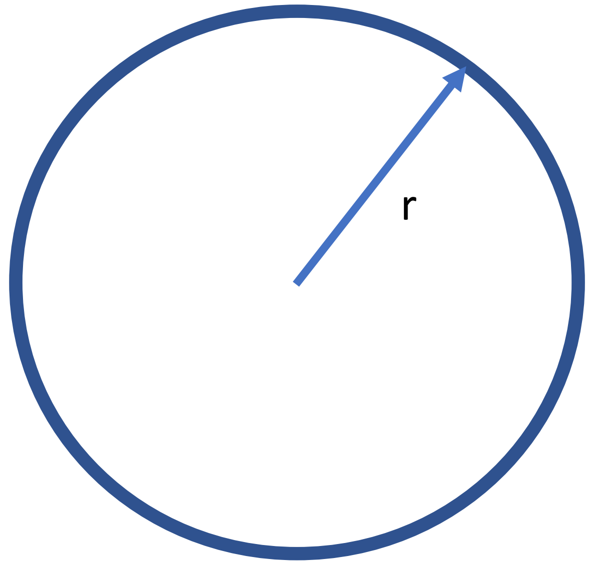
probability of false alarm

Idea: control the capacity of the descriptor space



Softmax: descriptors can fall anywhere

L2-Softmax: descriptors need to fall on the surface of a hypersphere

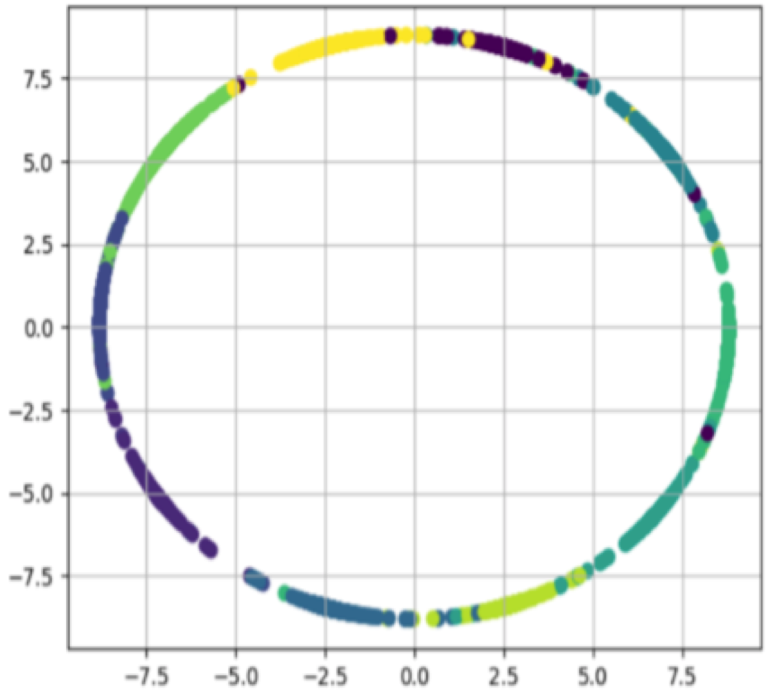


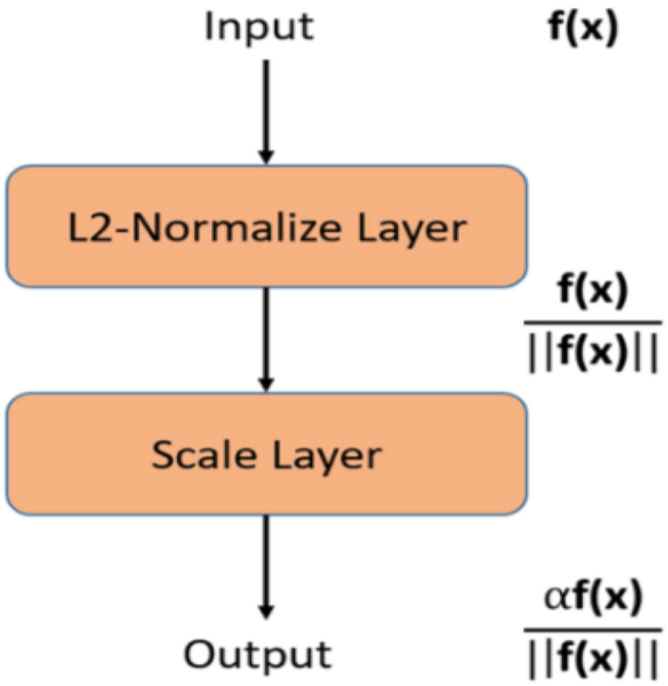
Rajeev Ranjan, Carlos D. Castillo, Rama Chellappa
L2-constrained softmax loss for discriminative face verification
Arxiv

Crystal Loss (L2 Softmax)

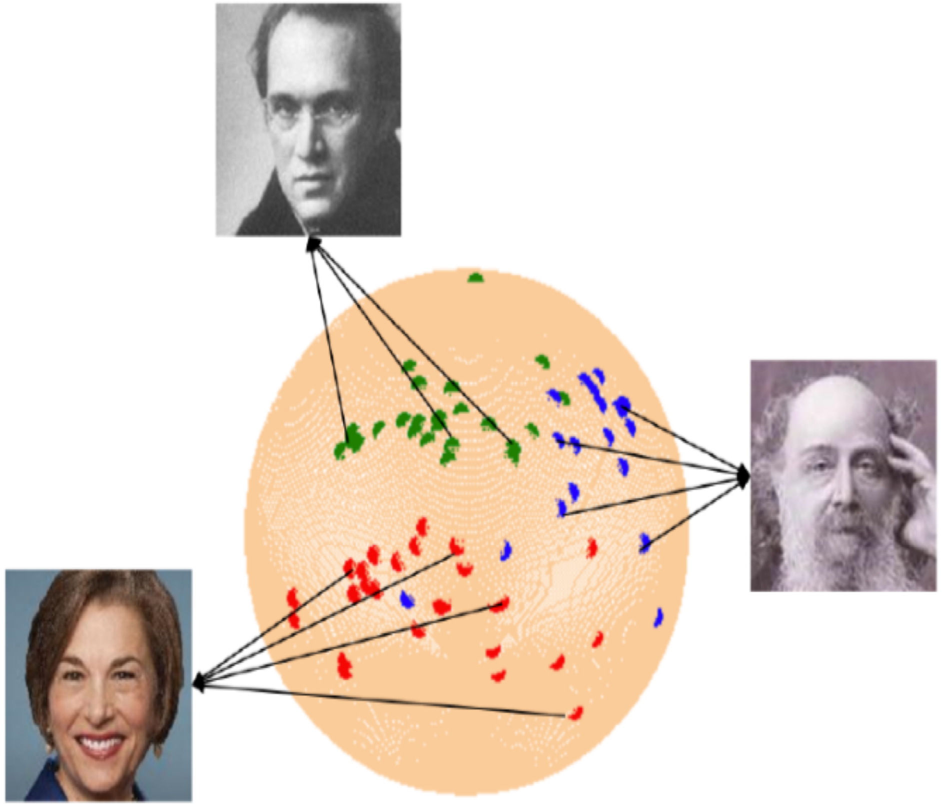
$$\text{minimize} \quad -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}$$

$$\text{subject to} \quad \|f(\mathbf{x}_i)\|_2 = \alpha, \ \ \forall i = 1, 2, \ldots M,$$

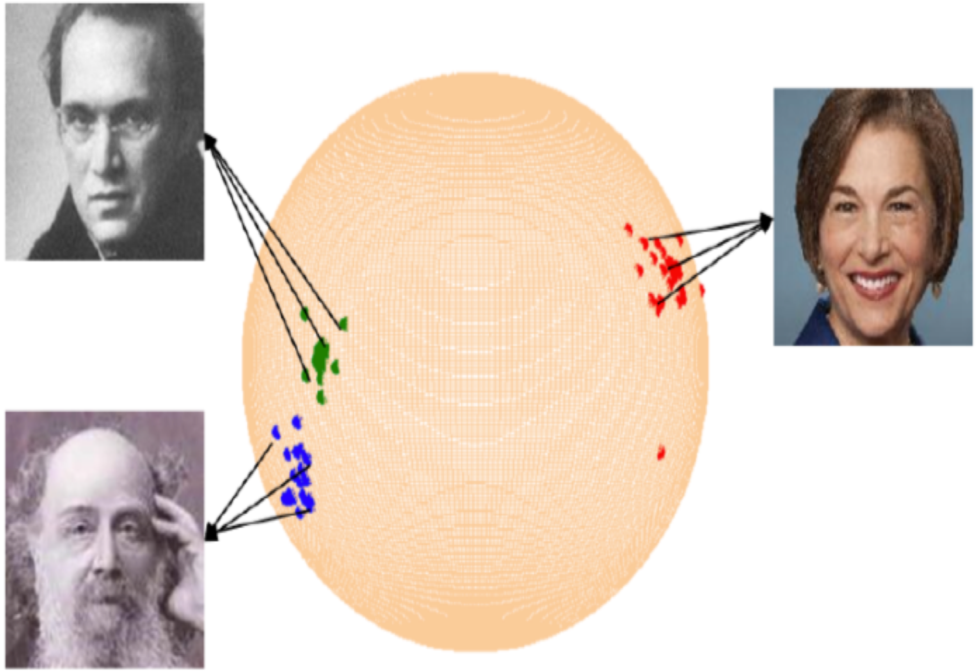




Embedding Quality



Softmax



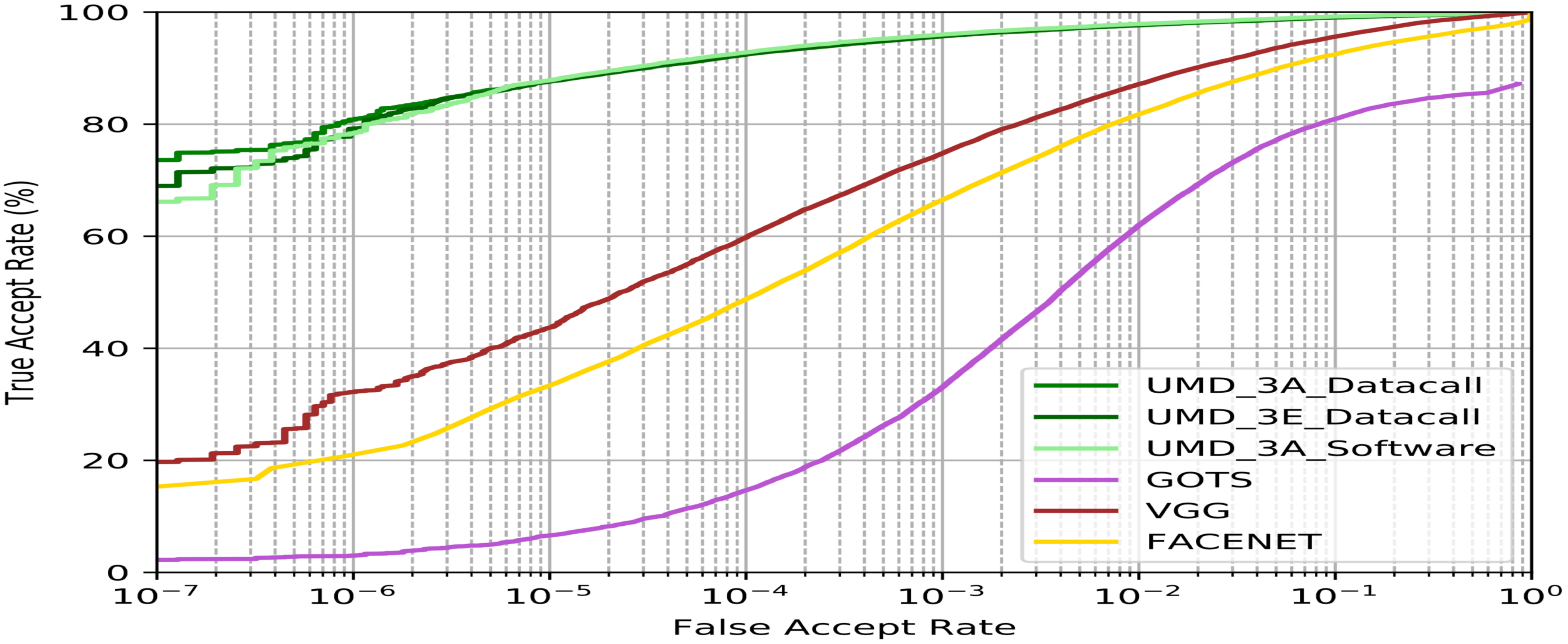
Crystal Loss

Results (Identification/ Verification IJB-A)

	IJB-A Verification (TAR @ FAR)			IJB-A Identification	
	0.0001	0.001	0.01	FPIR=0.01	Rank=1
VGG-Face	-	0.604	0.805	0.46	0.913
Chen et al.	-	-	0.838	-	0.903
Masi et al.	-	0.725	0.886	-	0.906
NAN	-	0.881	0.941	0.817	0.958
UltraFace	-	0.823	0.922	0.792	0.947
Crosswhite et al.	-	0.836	0.939	0.774	0.928
R101+L2S+ TPE	**0.898**	**0.942**	**0.969**	**0.910**	**0.971**
RX101+L2S+ TPE	**0.909**	**0.943**	**0.970**	**0.915**	**0.973**

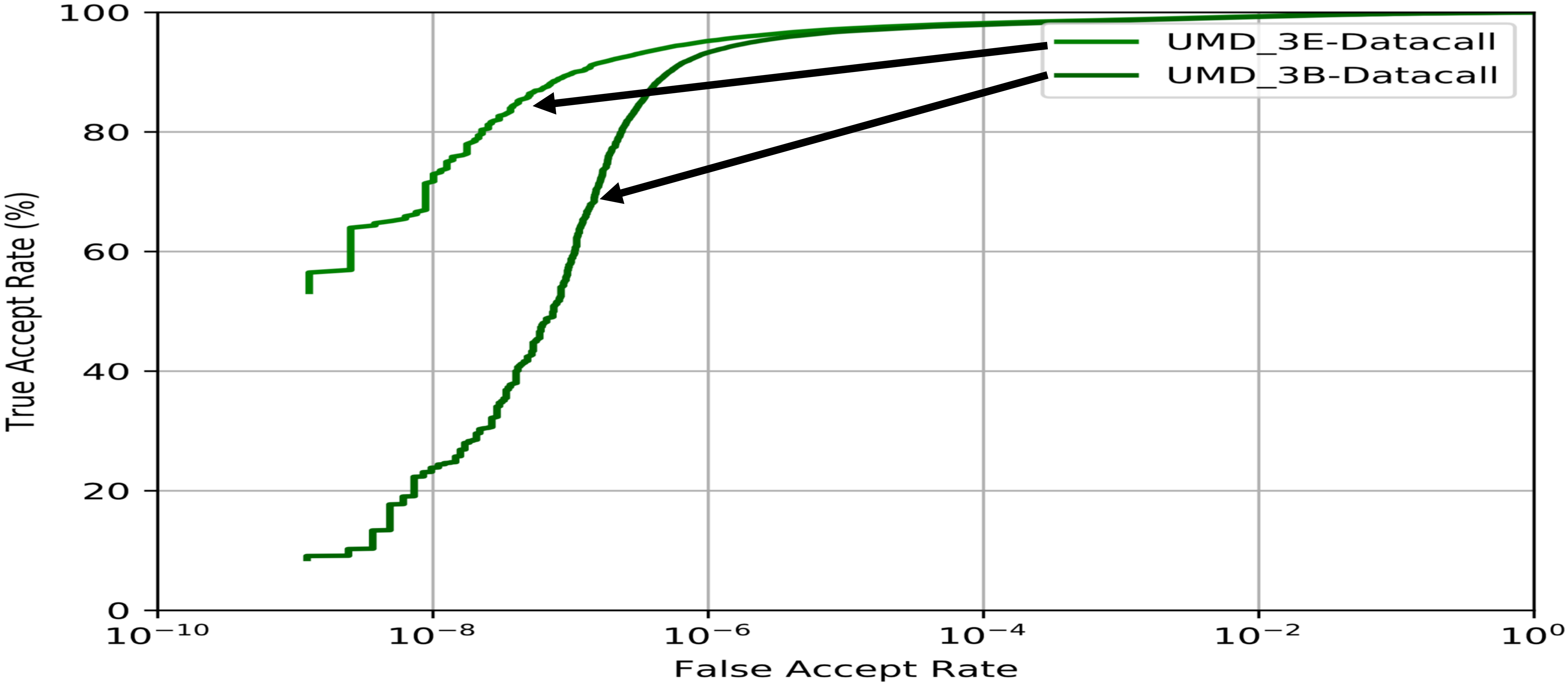
IJB-C Results with 1:1 Templates

recall



probability of false alarm

CS5 1:1, template to template, ROC



recall

probability of false alarm

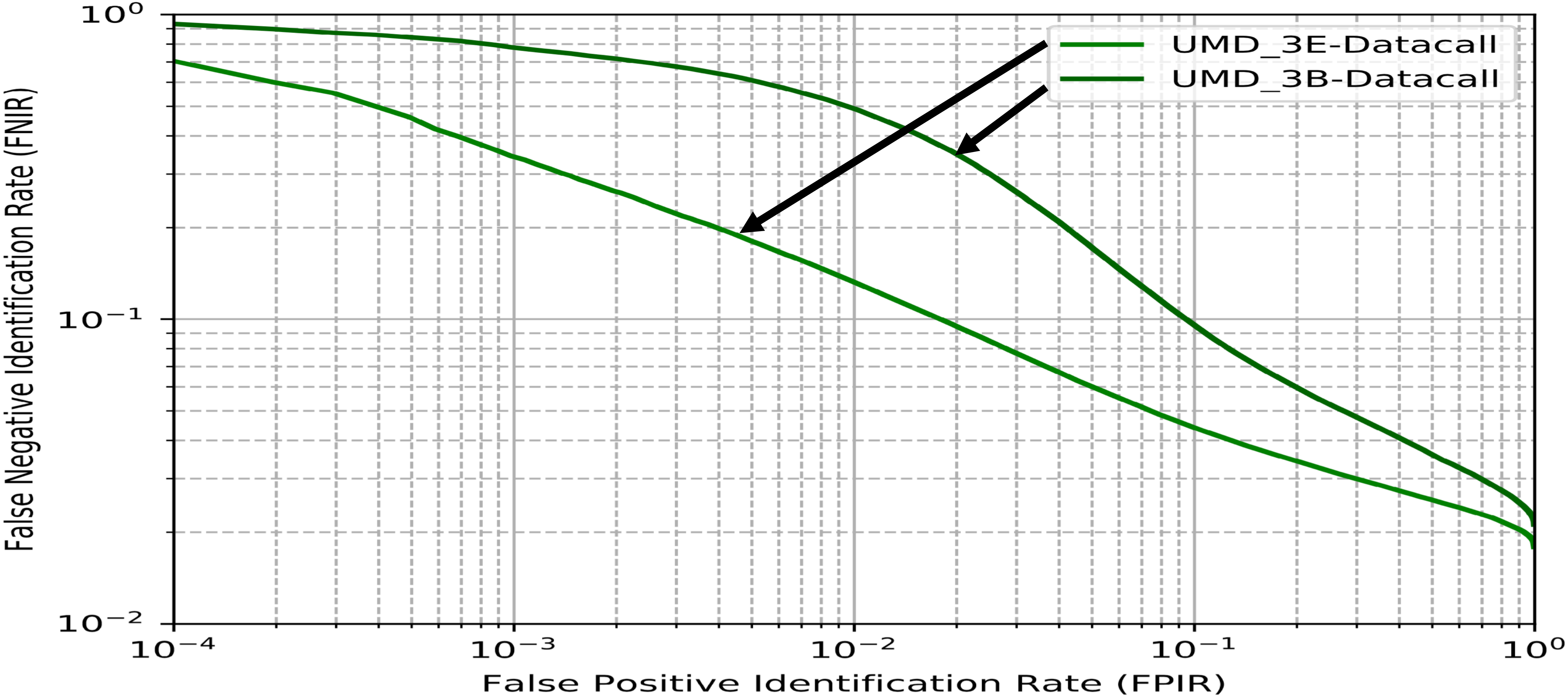
CS5 Search in a Gallery with 1M Individuals

- Define two galleries with a 1.1M individuals
- Perform 332K searches in each gallery
- Some searches have a matching item, some do not

	Accuracy % @ Rank										
	1	2	3	4	5	7	10	20	30	40	50
R-g1	96.62	97.21	97.40	97.51	97.59	97.68	97.78	97.96	98.05	98.12	98.16
RA_g1	**96.99**	**97.47**	**97.61**	**97.70**	**97.76**	**97.85**	**97.92**	**98.06**	**98.14**	**98.19**	**98.23**
RA_g1 (Index)	96.95	97.43	97.57	97.66	97.71	97.80	97.88	98.00	98.08	98.13	98.16

Open Set Metrics for CS5 Search

probability of rejecting a person enrolled in the gallery



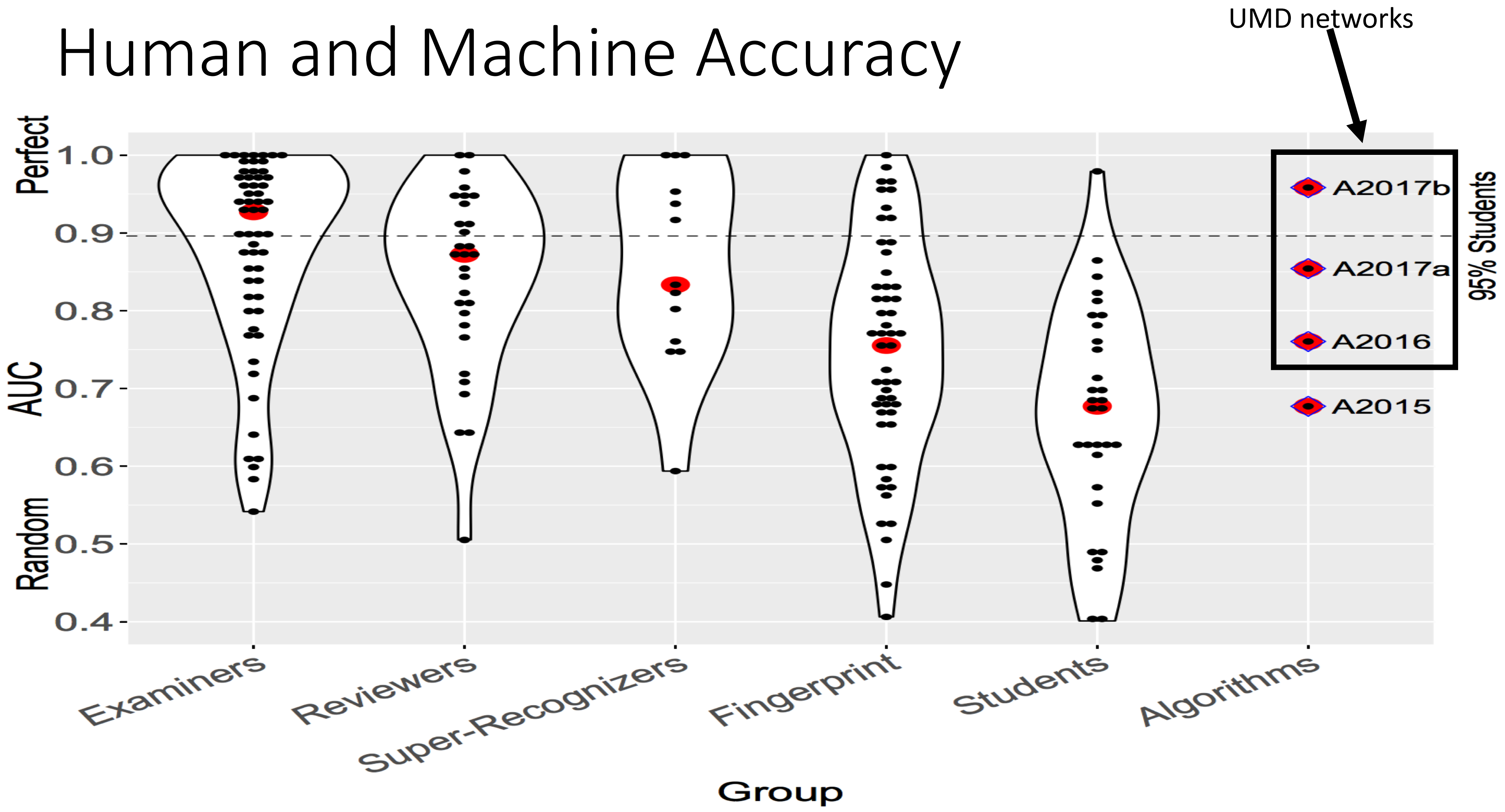
probability of accepting a person not enrolled in the gallery

Is this good compared to humans?

- We compared high-end systems like UMD’s to humans of different levels of abilities
  - Students, fingerprint examiners, super-recognizers, reviewers, examiners
- Will not say much, it has already been presented today by my coauthors, but I will give you my quick take.

P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Geraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O’Toole
Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms
PNAS 2018

Human and Machine Accuracy



Facial Recognition Experts Perform The Best With An AI Sidekick

CC0/Victor Tangermann

by Dan Robitzski | May 29, 2018 | Artificial Intelligence

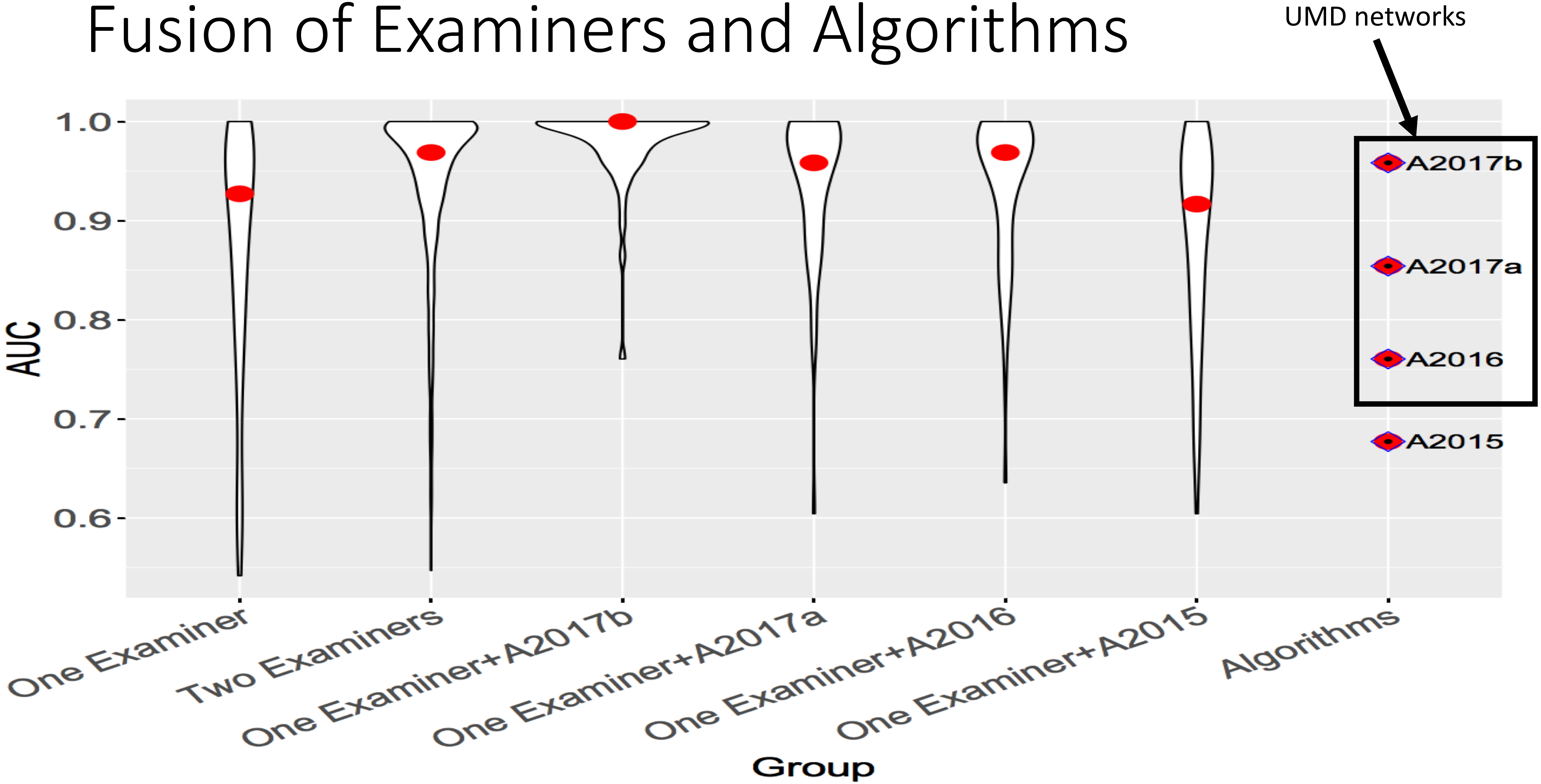
SHARE

Scientists are working on a kickass new twist to the classic buddy cop movie genre. Get this: cyberterrorist Marcus Hurricane is going to walk free unless police detective Rick Danger can place him at the scene of the crime. But all he has to go on are some grainy security camera images, and he can't quite make out Hurricane's signature badass face scars. Enter: detective Danger's trusty AI cyborg sidekick, Sparky. Together, they have what it takes to save the day.

#ai #artificial intelligence #facial recognition #forensic science

https://futurism.com/facial-recognition-experts-best-ai-sidekick/

Fusion of Examiners and Algorithms



Things we’ve Learned that are Important for High Performance in FR

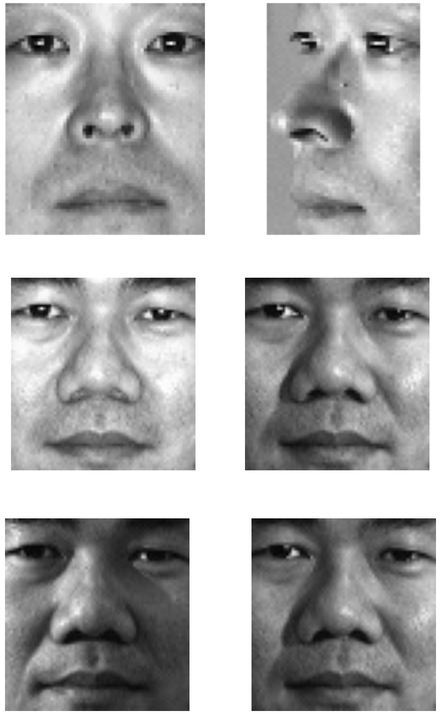
- Appearance is not identity and identity is not appearance
- Amount of data is key:
  - Number of images
  - Many identities
  - Carefully curated
- Loss functions
  - Being able to make sense of subtleties of face appearance/fine grained classification
  - Converging to useful solutions in datasets with many classes
- Alignment
  - Getting accurate key points throughout the pose continuum is still important, this enables accurate alignment.

Things we don’t know

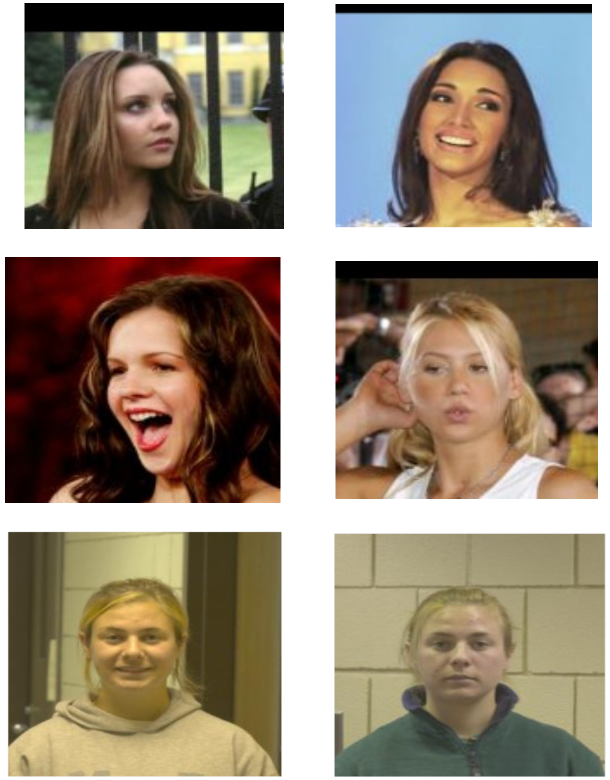
- Current state of the art methods for face recognition require:
  - (1) Lots of training data and (2) fully labeled data
  - How to handle the situation when we walk back those two requirements?
- How can computers and humans work together in verification and search tasks?
  - Are your reading glasses your sidekick?

The Future

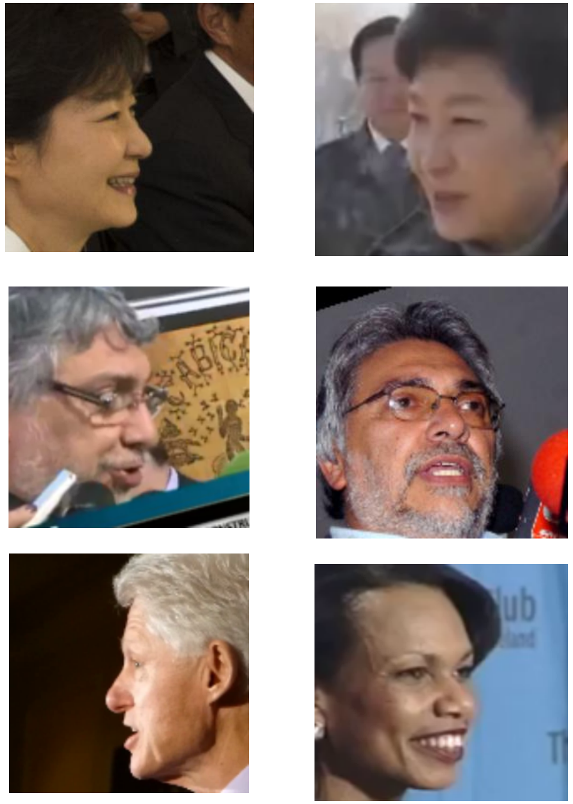
2006



2012



2018



2024

?